# s1061443_hw3

[1]
```
# s1061443_李杰穎
#先導入資料處理會用到的模組
import numpy as np
import numpy.random as random
import scipy as sp
from pandas import Series, DataFrame
import pandas as pd

# 可視化模組
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
%matplotlib inline

# 機器學習模組
import sklearn

# 表示到小數第三位
%precision 3
```

```
'%.3f'
```

[3]
```
# s1061443_李杰穎
data = pd.read_csv('student-por.csv')
data
```

| | school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | ... | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | ... | 4 | 3 | 4 | 1 | 1 | 3 | 4 | 0 | 11 | 11 |
| 1 | GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | ... | 5 | 3 | 3 | 1 | 1 | 3 | 2 | 9 | 11 | 11 |
| 2 | GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | ... | 4 | 3 | 2 | 2 | 3 | 3 | 6 | 12 | 13 | 12 |
| 3 | GP | F | 15 | U | GT3 | T | 4 | 2 | health | services | ... | 3 | 2 | 2 | 1 | 1 | 5 | 0 | 14 | 14 | 14 |
| 4 | GP | F | 16 | U | GT3 | T | 3 | 3 | other | other | ... | 4 | 3 | 2 | 1 | 2 | 5 | 0 | 11 | 13 | 13 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 644 | MS | F | 19 | R | GT3 | T | 2 | 3 | services | other | ... | 5 | 4 | 2 | 1 | 2 | 5 | 4 | 10 | 11 | 10 |
| 645 | MS | F | 18 | U | LE3 | T | 3 | 1 | teacher | services | ... | 4 | 3 | 4 | 1 | 1 | 1 | 4 | 15 | 15 | 16 |
| 646 | MS | F | 18 | U | GT3 | T | 1 | 1 | other | other | ... | 1 | 1 | 1 | 1 | 1 | 5 | 6 | 11 | 12 | 9 |
| 647 | MS | M | 17 | U | LE3 | T | 3 | 1 | services | services | ... | 2 | 4 | 5 | 3 | 4 | 2 | 6 | 10 | 10 | 10 |
| 648 | MS | M | 18 | R | LE3 | T | 3 | 2 | services | other | ... | 4 | 4 | 1 | 3 | 4 | 5 | 4 | 10 | 11 | 11 |

649 rows × 33 columns

[5]
```
# s1061443_李杰穎
from sklearn import linear_model
model = linear_model.LinearRegression()
```

[13]
```
# s1061443_李杰穎
X = data.loc[:, ['absences']].values
y = data['G3'].values
```

[14]
```
# s1061443_李杰穎
model.fit(X, y)
```

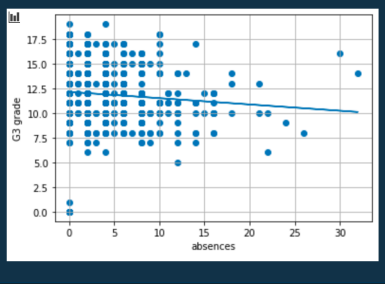
```
LinearRegression()
```

[16]
```
# s1061443_李杰穎
print('迴歸係數:', model.coef_)
print('截距:', model.intercept_)
print('決定係數:', model.score(X, y))
```

```
迴歸係數: [-0.064]
截距: 12.138800862687443
決定係數: 0.008350131955637385
```

[19]
```
# s1061443_李杰穎
plt.scatter(X, y)
plt.xlabel('absences')
plt.ylabel('G3 grade')

plt.plot(X, model.predict(X))
plt.grid(True)
```




[45]
```
# s1061443_李杰穎
# 計算各個行(欄位)裡有多少個"?"
auto = auto[['price', 'engine-size', 'width']]
auto.isin(['?']).sum()
```

```
price          0
engine-size    0
width          0
dtype: int64
```

[46]
```
# s1061443_李杰穎
# 將?取代為NaN，刪除有NaN的列
auto = auto.replace('?', np.nan).dropna()
print('汽車資料的形式:{}'.format(auto.shape))
```

```
汽車資料的形式:(201, 3)
```

[47]
```
# s1061443_李杰穎
#資料型別轉換
auto = auto.assign(price=pd.to_numeric(auto.price))
print('資料型別的確認 (型別轉換後) \n{}'.format(auto.dtypes))
```

```
資料型別的確認 (型別轉換後)
price            int64
engine-size      int64
width          float64
dtype: object
```

[48]
```
# s1061443_李杰穎
```

```python
# s1061443_李杰穎
# 為了資料分割(訓練資料與測試資料)的匯入
from sklearn.model_selection import train_test_split

# 為了多元線性迴歸模型建構的導入
from sklearn.linear_model import LinearRegression

# 指定目標變數為price、其他為解釋變數
X = auto.drop('price', axis=1)
y = auto['price']

# 分為訓練資料與測試資料
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=0)

# 多元線性迴歸的初始化學習
model = LinearRegression()
model.fit(X_train, y_train)

# 顯示決定係數
print('決定係數(train):{:3f}'.format(model.score(X_train, y_train)))
print('決定係數(test):{:3f}'.format(model.score(X_test, y_test)))
# 顯示迴歸係數與截距
print('\n迴歸係數\n{}'.format(pd.Series(model.coef_, index=X.columns)))
print('截距: {:.3f}'.format(model.intercept_))
```

```
決定係數(train):0.783189
決定係數(test):0.778292

迴歸係數
engine-size     109.526787
width          1261.735518
dtype: float64
截距: -84060.643
```